Assignment 4 MapReduce

Task 1: Write a Map-Reduce Program to filter invalid record.

Text file


```
[acadgild@localhost mapreduce]$ cat television.txt
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Akai|Decent|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
Onida|Lucid|18|Uttar Pradesh|232401|16200
Onida|Decent|14|Uttar Pradesh|232401|16200
Onida|NA|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Zen|Super|14|Maharashtra|619082|9200
Samsung|Optima|14|Madhya Pradesh|132401|14200
NA|Lucid|18|Uttar Pradesh|232401|16200
Samsung|Decent|16|Kerala|922401|12200
Lava|Attention|20|Assam|454601|24200
Samsung|Super|14|Maharashtra|619082|9200
Samsung|Super|14|Maharashtra|619082|9200
Samsung|Super|14|Maharashtra|619082|9200[acadgild@localhost mapreduce]$
```

Command:

hadoop jar assign4.jar mrTask1.Task1 /mapreduce/television.txt /mapreduce/task1output

Code Snipet:


```
[acadgild@localhost mapreduce]$ hadoop jar assign4.jar mrTask1.Task1  /mapreduce/television.txt /mapreduce/task1output
19/01/11 06:09:14 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/01/11 06:09:15 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
19/01/11 06:09:17 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application wit
ToolRunner to remedy this.
19/01/11 06:09:17 INFO input.FileInputFormat: Total input paths to process : 1
19/01/11 06:09:17 INFO mapreduce.JobSubmitter: number of splits:1
19/01/11 06:09:18 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1547166974239_0001
19/01/11 06:09:18 INFO impl.YarnClientImpl: Submitted application application_1547166974239_0001
19/01/11 06:09:18 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1547166974239_0001/
19/01/11 06:09:18 INFO mapreduce.Job: Running job: job_1547166974239_0001
19/01/11 06:09:32 INFO mapreduce.Job: Job job_1547166974239_0001 running in uber mode : false
19/01/11 06:09:32 INFO mapreduce.Job:  map 0% reduce 0%
19/01/11 06:09:39 INFO mapreduce.Job:  map 100% reduce 0%
19/01/11 06:09:39 INFO mapreduce.Job: Job job_1547166974239_0001 completed successfully
19/01/11 06:09:39 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=107363
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=844
                HDFS: Number of bytes written=73
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=4850
                Total time spent by all reduces in occupied slots (ms)=0
```

```
19/01/11 06:09:32 INFO mapreduce.Job: Job job_1547166974239_0001 running in uber mode : false
19/01/11 06:09:32 INFO mapreduce.Job:  map 0% reduce 0%
19/01/11 06:09:39 INFO mapreduce.Job:  map 100% reduce 0%
19/01/11 06:09:39 INFO mapreduce.Job: Job job_1547166974239_0001 completed successfully
19/01/11 06:09:39 INFO mapreduce.Job: Counters: 30
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=107363
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=844
                HDFS: Number of bytes written=73
                HDFS: Number of read operations=5
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=4850
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=4850
                Total vcore-milliseconds taken by all map tasks=4850
                Total megabyte-milliseconds taken by all map tasks=4966400
        Map-Reduce Framework
                Map input records=18
                Map output records=2
                Input split bytes=111
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=66
                CPU time spent (ms)=510
                Physical memory (bytes) snapshot=106995712
                Virtual memory (bytes) snapshot=2056757248
                Total committed heap usage (bytes)=62980096
        File Input Format Counters
                Bytes Read=733
        File Output Format Counters
                Bytes Written=73
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost mapreduce]$ hadoop jar assign4.jar mrTask2.Task2  /mapreduce/television.txt /mapreduce/task2output
```

Output Snipet:


```
[acadgild@localhost ~]$ hdfs dfs -cat /mapreduce/task
19/01/11 06:10:39 WARN util.NativeCodeLoader: Unable
        Onida|NA|16|Kerala|922401|12200
        NA|Lucid|18|Uttar Pradesh|232401|16200
[acadgild@localhost ~]$
```

Task 2: Write a Map-Reduce Program to calculate the total unit sold for each company

Command:

hadoop jar assign4.jar mrTask2.Task2 /mapreduce/television.txt /mapreduce/task2output

Code Snipet:

```
[acadgild@localhost mapreduce]$ hadoop jar assign4.jar mrTask2.Task2  /mapreduce/television.txt /mapreduce/task2output
19/01/11 06:17:03 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/01/11 06:17:04 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
19/01/11 06:17:05 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application with
ToolRunner to remedy this.
19/01/11 06:17:06 INFO input.FileInputFormat: Total input paths to process : 1
19/01/11 06:17:06 INFO mapreduce.JobSubmitter: number of splits:1
19/01/11 06:17:06 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1547166974239_0002
19/01/11 06:17:06 INFO impl.YarnClientImpl: Submitted application application_1547166974239_0002
19/01/11 06:17:07 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1547166974239_0002/
19/01/11 06:17:07 INFO mapreduce.Job: Running job: job_1547166974239_0002
19/01/11 06:17:16 INFO mapreduce.Job: Job job_1547166974239_0002 running in uber mode : false
19/01/11 06:17:16 INFO mapreduce.Job:  map 0% reduce 0%
19/01/11 06:17:22 INFO mapreduce.Job:  map 100% reduce 0%
19/01/11 06:17:32 INFO mapreduce.Job:  map 100% reduce 100%
19/01/11 06:17:32 INFO mapreduce.Job: Job job_1547166974239_0002 completed successfully
19/01/11 06:17:33 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=225
                FILE: Number of bytes written=215827
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=844
                HDFS: Number of bytes written=43
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=4869
                Total time spent by all reduces in occupied slots (ms)=6575
                Total time spent by all map tasks (ms)=4869
                Total time spent by all reduce tasks (ms)=6575
                Total vcore-milliseconds taken by all map tasks=4869
                Total vcore-milliseconds taken by all reduce tasks=6575
                Total megabyte-milliseconds taken by all map tasks=4985856
                Total megabyte-milliseconds taken by all reduce tasks=6732800
        Map-Reduce Framework
                Map input records=18
```

```
                Total vcore-milliseconds taken by all reduce tasks=6575
                Total megabyte-milliseconds taken by all map tasks=4985856
                Total megabyte-milliseconds taken by all reduce tasks=6732800
        Map-Reduce Framework
                Map input records=18
                Map output records=18
                Map output bytes=183
                Map output materialized bytes=225
                Input split bytes=111
                Combine input records=0
                Combine output records=0
                Reduce input groups=6
                Reduce shuffle bytes=225
                Reduce input records=18
                Reduce output records=6
                Spilled Records=36
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=196
                CPU time spent (ms)=1530
                Physical memory (bytes) snapshot=320380928
                Virtual memory (bytes) snapshot=4118196224
                Total committed heap usage (bytes)=222429184
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=733
        File Output Format Counters
                Bytes Written=43
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost mapreduce]$
```

Output Snipet:

```
[acadgild@localhost ~]$ hdfs dfs -cat /mapreduce/task2output/part-r-00000
19/01/11 06:20:08 WARN util.NativeCodeLoader: Unable to load native-hadoop
Akai     1
Lava     3
NA       1
Onida    4
Samsung  7
Zen      2
You have new mail in /var/spool/mail/acadgild
```

Task 3: Write a Map-Reduce Program to calculate the total unit sold in each state for Onida company.

Command:

hadoop jar assign4.jar mrTask3.Task3 /mapreduce/television.txt /mapreduce/task3output

Code Snipet:

```
[acadgild@localhost mapreduce]$ hadoop jar assign4.jar mrTask3.Task3  /mapreduce/television.txt /mapreduce/task3output
19/01/11 06:20:54 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
19/01/11 06:20:55 INFO client.RMProxy: Connecting to ResourceManager at localhost/127.0.0.1:8032
19/01/11 06:20:56 WARN mapreduce.JobResourceUploader: Hadoop command-line option parsing not performed. Implement the Tool interface and execute your application wi
ToolRunner to remedy this.
19/01/11 06:20:57 INFO input.FileInputFormat: Total input paths to process : 1
19/01/11 06:20:57 INFO mapreduce.JobSubmitter: number of splits:1
19/01/11 06:20:57 INFO mapreduce.JobSubmitter: Submitting tokens for job: job_1547166974239_0003
19/01/11 06:20:57 INFO impl.YarnClientImpl: Submitted application application_1547166974239_0003
19/01/11 06:20:57 INFO mapreduce.Job: The url to track the job: http://localhost:8088/proxy/application_1547166974239_0003/
19/01/11 06:20:57 INFO mapreduce.Job: Running job: job_1547166974239_0003
19/01/11 06:21:07 INFO mapreduce.Job: Job job_1547166974239_0003 running in uber mode : false
19/01/11 06:21:07 INFO mapreduce.Job:  map 0% reduce 0%
19/01/11 06:21:14 INFO mapreduce.Job:  map 100% reduce 0%
19/01/11 06:21:21 INFO mapreduce.Job:  map 100% reduce 100%
19/01/11 06:21:21 INFO mapreduce.Job: Job job_1547166974239_0003 completed successfully
19/01/11 06:21:21 INFO mapreduce.Job: Counters: 49
        File System Counters
                FILE: Number of bytes read=387
                FILE: Number of bytes written=216137
                FILE: Number of read operations=0
                FILE: Number of large read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=844
                HDFS: Number of bytes written=64
                HDFS: Number of read operations=6
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=4582
                Total time spent by all reduces in occupied slots (ms)=4945
                Total time spent by all map tasks (ms)=4582
                Total time spent by all reduce tasks (ms)=4945
                Total vcore-milliseconds taken by all map tasks=4582
                Total vcore-milliseconds taken by all reduce tasks=4945
                Total megabyte-milliseconds taken by all map tasks=4691968
                Total megabyte-milliseconds taken by all reduce tasks=5063680
        Map-Reduce Framework
                Map input records=18
                Map output records=22
```

```
                Launched reduce tasks=1
                Data-local map tasks=1
                Total time spent by all maps in occupied slots (ms)=4582
                Total time spent by all reduces in occupied slots (ms)=4945
                Total time spent by all map tasks (ms)=4582
                Total time spent by all reduce tasks (ms)=4945
                Total vcore-milliseconds taken by all map tasks=4582
                Total vcore-milliseconds taken by all reduce tasks=4945
                Total megabyte-milliseconds taken by all map tasks=4691968
                Total megabyte-milliseconds taken by all reduce tasks=5063680
        Map-Reduce Framework
                Map input records=18
                Map output records=22
                Map output bytes=337
                Map output materialized bytes=387
                Input split bytes=111
                Combine input records=0
                Combine output records=0
                Reduce input groups=5
                Reduce shuffle bytes=387
                Reduce input records=22
                Reduce output records=5
                Spilled Records=44
                Shuffled Maps =1
                Failed Shuffles=0
                Merged Map outputs=1
                GC time elapsed (ms)=157
                CPU time spent (ms)=1320
                Physical memory (bytes) snapshot=320028672
                Virtual memory (bytes) snapshot=4118196224
                Total committed heap usage (bytes)=222429184
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
        File Input Format Counters
                Bytes Read=733
        File Output Format Counters
                Bytes Written=64
[acadgild@localhost mapreduce]$
```

Output Snipet:

```
[acadgild@localhost ~]$ hdfs dfs -cat /mapreduce/task3output/part-r-00000
19/01/11 06:26:54 WARN util.NativeCodeLoader: Unable to load native-hadoop libr
Assam    0
Kerala   1
Madhya Pradesh   0
Maharashtra      0
Uttar Pradesh    3
You have new mail in /var/spool/mail/acadgild
[acadgild@localhost ~]$
```